# Users and Assessors in the Context of INEX: Are Relevance Dimensions Relevant?

Jovan Pehcevski, James A. Thom, Anne-Marie Vercoustre

# Users and Assessors in the Context of INEX: Are Relevance Dimensions Relevant?

Jovan Pehcevski
School of CS and IT
RMIT University
Melbourne, Australia
jovanp@cs.rmit.edu.au

James A. Thom
School of CS and IT
RMIT University
Melbourne, Australia
jat@cs.rmit.edu.au

Anne-Marie Vercoustre
AxIS research group
INRIA
Rocquencourt, France
anne-marie.vercoustre@inria.fr



## ABSTRACT

*The main aspects of XML retrieval are identified by analysing and comparing the following two behaviours: the behaviour of the assessor when judging the relevance of returned document components; and the behaviour of users when interacting with components of XML documents. We argue that the two INEX relevance dimensions, Exhaustivity and Specificity, are not orthogonal dimensions; indeed, an empirical analysis of each dimension reveals that the grades of the two dimensions are correlated to each other. By analysing the level of agreement between the assessor and the users, we aim at identifying the best units of retrieval. The results of our analysis show that the highest level of agreement is on highly relevant and on non-relevant document components, suggesting that only the end points of the INEX 10-point relevance scale are perceived in the same way by both the assessor and the users. We propose a new definition of relevance for XML retrieval and argue that its corresponding relevance scale would be a better choice for INEX.*

## 1. INTRODUCTION

The INitiative for the Evaluation of XML retrieval[1] (INEX) is a coordinated effort that promotes evaluation procedures for content-oriented XML retrieval. In order to evaluate XML retrieval effectiveness, the concept of *relevance* needs to be clearly defined. There are two *relevance dimensions* used by INEX, *Exhaustivity* and *Specificity*, which measure the extent to which a given information unit *covers* and is *focused on* an information need, respectively [16]. In this paper we provide a detailed empirical analysis of the two INEX relevance dimensions. More specifically, we investigate what the experience of both the assessor and the users suggests on how relevance should be defined and measured in the context of XML retrieval.

The INEX test collection consists of three parts: an XML document collection, a set of topics required to search for information stored in this collection, and a set of relevance assessments that correspond to these topics [12]. The XML document collection comprises 12,107 IEEE Computer Society articles published in the period between 1997-2002, with approximately 500MB of data. To search for information stored in this collection, two types of topics are explored in INEX: Content-Only (CO) topics and Content-And-Structure (CAS) topics. CO topics do not refer to the existing document structure, whereas CAS topics enforce restrictions on the document structure and explicitly specify the target element. In this paper, we focus on the CO topics to analyse the behaviour of the assessor and the users in the context of INEX.

Tombros et al. [20] demonstrate that, while assessing relevance of retrieved pages on the Web, the context determined by a task type has an effect on the user behaviour. A similar effect is likely to be expected when users assess the relevance of XML document components (rather than of whole documents, such as Web pages) [19]. The CO topics used in this study are thus selected such that they correspond to different types of tasks, or different *topic categories*: a *Background* category and a *Comparison* category.

Since 2002, a new set of topics has been introduced and assessed by INEX participants each year. Analysing the behaviour of assessors when judging the relevance of returned document components may provide insight into the possible trends within the relevance judgements. Such studies have been done for both the INEX 2002 [9] and the INEX 2003 [16] test collections. We have recently also analysed the relevance judgements of the INEX 2004 topics, where we aimed at understanding what assessors consider to be the most useful answers [14].

There is growing interest among the research community in studying the user behaviour in the context of XML retrieval; however, little work has been done in the field so far. The most notable is the work done by Finesilver and Reid [4], where a small-scale experimental study is designed to investigate the information-seeking behaviour of users in the context of structured documents. Recently, an Interactive track was established at INEX 2004 to investigate the retrieval behaviour of users when components of XML documents – estimated as likely to be relevant by an XML retrieval system – are presented as answers [19]. Ten of the 43 active research groups in INEX 2004 were also involved in the Interactive track, and each group was required to provide a minimum of eight users to interact with the retrieval system. The analysis of the user behaviour in this paper is based on the user judgements provided by these groups.

When judging the relevance of a document component, two relevance dimensions – *Exhaustivity* and *Specificity* – are used by INEX. Each dimension uses four grades of relevance.

[1] `http://inex.is.informatik.uni-duisburg.de/2005/`

To assign a relevance score to a document component, the grades from each dimension are combined into a single 10-point relevance scale. However, the latter choice of combining the grades poses the following question: *is the 10-point relevance scale well perceived by users*?

Due to hierarchical relationships between the XML document components, an XML retrieval system may often return components with varying granularity. The problem that often arises in this retrieval scenario is the one of distinguishing the *appropriate level of retrieval granularity*. This problem, which is often referred to as the *overlap problem*, remains an open research problem in the field of XML retrieval. Indeed, it has been shown that it is not only a retrieval problem [14, 15], but also a serious evaluation problem [8]. This then raises the question: *is retrieving overlapping document components what users really want*?

In this work, we aim at finding answers to the above research questions. We show that the overlap problem is handled differently by the assessor and the users, and that the two INEX relevance dimensions are perceived as one. We propose a new definition of relevance for INEX and argue that its corresponding relevance scale would bring a better value for the XML retrieval evaluation.

The remainder of the paper is organised as follows. In Section 2 we provide an overview of the methodology used in this study. The concept of *relevance* in information retrieval is thoroughly discussed in Section 3, where we particularly focus on how the INEX definition of relevance fits in the unified relevance framework. We study the behaviour of the assessor and the users in Sections 4 and 5, when two categories of retrieval topics are considered, respectively. Our new definition of relevance is described in Section 6. We conclude in Section 7 with a brief discussion of our findings.

## 2. METHODOLOGY

In this section, we provide a detailed overview of the methodology used in this study. More precisely, we describe the type and the number of participants involved; the choice of the two categories of topics used; and the way the data – reflecting the observed behaviour of participants – was collected. The data reflecting the observed behaviour, as analysed in this study, was collected from well-established INEX activities, which are also explained in separate studies. For instance, for a particular CO topic we use the relevance judgements obtained from the interactive online assessment system [16] to analyse the behaviour of the assessor. Similarly, for the same topic we use the data collected for the purposes of the INEX 2004 Interactive track [19] to analyse the retrieval behaviour of users. We actively participated in both INEX activities.

### 2.1 Participants

Two types of participants are used in this study: assessors and users. In general, both can be regarded as users; however, it is often necessary to distinguish between them, since their purpose in the XML retrieval task is quite different.

#### *Assessors*

Every year since 2002 when INEX started, each participant is asked to submit at least one retrieval topic (query). If a topic is accepted, the same participant is (usually) required to assess the relevance of the retrieved document components. The assessor can, therefore, be seen as an entity that provides the ground-truth for a particular retrieval topic. There is usually one assessor per topic, although for the purpose of checking whether the relevance judgements were done in a consistent manner, two or more assessors may be assigned to a given topic [16]. In this study we analyse the relevance assessments provided by one assessor per topic.

```
Topic B1 (INEX 2004 CO topic 192):

You are writing a large article discussing virtual reality
(VR) applications and you need to discuss their negative
side effects. What you want to know is the symptoms
associated with cybersickness, the amount of users who get
them, and the VR situations where they occur. You are not
interested in the use of VR in therapeutic treatments
unless they discuss VR side effects.
```

**Figure 1: A *Background* topic example.**

#### *Users*

A total of 88 users were employed for the purposes of the Interactive track at INEX 2004, with an average age of 29 years [19]. Although most of the users had a substantial level of experience in Web or other related searches, it was expected that very few (if any) were experienced in interacting with XML document components. For this purpose, users were given the same (or rather, slightly modified) retrieval topics as the ones proposed and judged by the assessors. Analysing the data collected from the user interaction may thus indicate how well an XML retrieval system succeeds in satisfying users' information needs. Our analysis in this study is based on the user judgements provided by roughly 50 users per topic.

### 2.2 Retrieval Topics

To make users better understand the objectives of the retrieval task, the CO topics were reformulated as simulated work task situations [19]. A simulated work task situation requires users to interact with the retrieval system, which in turn – by allowing users to formulate as many queries as needed – results in different individual interpretations of the information need [2]. Thus, the reformulated CO topics not only describe *what* the information need represents, but also *why* users need to satisfy this need, and what is the *context* where the information need arises.

The CO topics used in the INEX Interactive track are divided in two task categories: a *Background* category and a *Comparison* category. Topics that belong to the *Background* category seek to find as much general information about the area of interest as possible. Two retrieval topics were used in this category, B1 and B2, which are based on the INEX 2004 CO topics 192 and 180, respectively [19]. Figure 1 shows Topic B1, which is the *Background* topic used in this study. Topics that belong to the *Comparison* category seek to find similarities or differences between at least two items discussed in the topic. Two retrieval topics were used in this category, C1 and C2, which are respectively based on

```
Topic C2 (INEX 2004 CO topic 198):

You are working on a project to develop a next generation
version of a software system. You are trying to decide on
the benefits and problems of implementation in a number
of programming languages, but particularly Java and Python.
You would like a good comparison of these for application
development. You would like to see comparisons of Python
and Java for developing large applications. You want to
see articles, or parts of articles, that discuss the
positive and negative aspects of the languages. Things
that discuss either language with respect to application
development may be also partially useful to you. Ideally,
you would be looking for items that are discussing both
efficiency of development and efficiency of execution
time for applications.
```

**Figure 2: A *Comparison* topic example.**

the INEX 2004 CO topics 188 and 198 [19]. Figure 2 shows Topic C2, which is the *Comparison* topic used in this study.

The motivation of using topics B1 and C2 in our study comes from the fact that both of these topics have corresponding relevance judgements available, and that data from roughly 50 users was collected for each of these topics. In contrast, no relevance judgements are available for topic B2, while data from around 18 users was collected for each of the topics B2 and C1. Previous work has also shown that XML retrieval systems exhibit varying behaviour when their performance is evaluated against different CO topic categories [7, 15]. It is then reasonable to expect that the level of agreement between the assessor and the users, which concerns the choice of the best units of retrieval, may depend on the topic category. Thus, in our forthcoming analysis of the retrieval behaviour, we clearly distinguish between topics B1 and C2.

## 2.3 Collecting the Data

Different means were used to collect the data from the assessor and the users, and different time restrictions were put in place in both cases.

In the case of the assessor, an interactive online assessment system is used to collect the judgements for a particular topic [16]. This is a well-established method used in INEX, where the assessment system implements some rules to ensure that the collected relevance judgements are as exhaustive and as consistent as possible. On average it takes one week for the assessor to judge all the retrieved elements for a particular topic. The relevance judgements are then stored in an *XML assessment file* where, for each XML document retrieved by participant systems, the judged elements are kept in document order. We use two assessment files, one for each topic B1 and C2, to analyse the relevance judgements made by assessors.

For users, a system based on HyREX [6] is used to collect the user judgements and to log their activities. Tombros et al. [19] explain the process of user interaction with the HyREX system in detail. Users are able to choose between two retrieval topics for each topic category, for which they are required to find as much information as possible for completing the search task. A time limit of 30 minutes is given to each user. The data obtained from the user interaction is stored in corresponding log files. For each user, we create an assessment file that follows the same structure as the assessor's assessment file. We use these files to analyse the judgements made by users for each of the topics B1 and C2.

An important point to note is that HyREX uses the concept of "index objects" [6] to limit the level of retrieval granularity that will be returned to users. This means that users were able to make judgements for only four (out of 192) element names. These names are `article`, `sec`, `ss1`, and `ss2`, which correspond to full article and to section and subsection elements of varying nesting levels, respectively. Although this may be seen as a limitation of the HyREX system, the obtained element granularity is nevertheless sufficient for the purpose of our analysis. To be consistent in our comparison of the observed behaviour between the assessor and the users, all element names different from these four were also removed from the two files containing assessors' judgements. If an element has been judged more than once, either by a user or an assessor, only the last relevance judgement is stored in the assessment files.

## 2.4 Measuring Overlap

When collecting assessor or user judgements for a particular topic, we also measure the level of overlap between the judged elements. There are *at least* two ways by which the overlap can be measured:

- *set-based overlap*, which for a *set* of returned elements measures the percentage of elements for which there exists another element that *fully contains* them; and
- *list-based overlap*, which takes into account the order of processing of returned elements, and measures the percentage of elements for which there exists another element *higher in the list* that fully contains them.

Consider the following set of returned elements:

1. `/article[1]/sec[1]`
2. `/article[1]/sec[1]/ss1[1]`
3. `/article[1]/sec[1]/ss1[1]/ss2[1]`
4. `/article[1]/sec[2]/ss1[1]`
5. `/article[1]/sec[2]`

Let us assume that the elements are returned in the above order, and that all the elements belong to one XML document. The set-based overlap in this case would be 60%, because three (out of five) elements in this set are fully contained by other element in the set (the three elements are the ones belonging to ranks 2, 3 and 4). The list-based overlap, however, would be 40%, because there are only two elements for which there exists another element higher in the list that fully contains them (the two elements that belong to ranks 2 and 3).

In this study we use the set-based overlap, as defined above, to measure the overlap between the judged elements. How-

ever, unlike in the assessor's case where the relevance judgements were obtained from only one assessor, the user judgements for a given topic were obtained from more than one user. To deal with this issue in a consistent manner, in users' case we measure the overlap *separately for each user*, and take the average to represent the resulting set-based overlap.

# 3. RELEVANCE: DEFINITIONS AND DIMENSIONS

It is a commonly held view that *relevance* is one of the most important concepts for the fields of documentation, information science, and information retrieval [13, 17]. Indeed, the main purpose of a retrieval system is to retrieve units of information estimated as *likely to be relevant* to an information need, as represented by a query. To build and evaluate effective information retrieval systems, the concept of relevance needs to be clearly defined and formalised.

Mizzaro [13] provides an overview of different definitions of relevance. These are also conveniently summarised by Lavrenko [10]. In general, there is a system-oriented, a user-oriented, and a logical definition of relevance. However, there are also other definitions of relevance, which relate to its nature and the notion of dependence. With respect to its nature, there is a binary or non-binary (graded) relevance. With respect to whether the relevance of a retrieved unit is dependent or not on any other unit already inspected by the user, there is a dependent or independent relevance. In the case of the former, the relevance is often distinguished either as a relevance conditional to a set of relevant retrieved units, or as a novel relevance, or as an aspect relevance.

In the following we provide an overview of several definitions of relevance, including the INEX relevance definition. We then describe a notable attempt to construct a unified definition of relevance [13].

## 3.1 System-oriented Relevance Definition

The system-oriented definition provides a binary relation between a unit of information (a document or a document component) and a user request (a query). To model this relation, both the unit of information and the user request are represented by a set of terms, reflecting the contents of the unit and the interest of the user, respectively. In this case, relevance is simply defined by the level of semantic overlap between the two representations; the more similar these representations are, the more likely the information unit is relevant to the user request. According to this definition, relevance is not dependent on any factors other than the two representations above. More precisely, it depends neither on the user who issued the request (or on the user information need, for that matter), nor on any other information units (regardless of whether they have been previously considered to be relevant or not), nor on any other requests to which the unit of information may or may not be relevant.

## 3.2 Novel Relevance

Novel relevance deals with the impact of retrieving redundant information units on user's perception of relevance. For example, if a system retrieves two near-duplicate information units, which may both be relevant to a request, the user will very likely not be interested in reading both of them, since once the first one is read, the second becomes entirely redundant. Carbonell and Goldstein proposed the concept of *Maximal Marginal Relevance* [3], which attempts to provide a balance between the relevance of a document to a query, and the redundancy of that document with respect to all the other documents previously inspected by the user. An interesting approach that may be seen as an extension of the above work was proposed by Allan et al. [1]. Their work attempts to address redundancy on a sub-document level and is based on the following idea: even if a document is considered to be mostly redundant by a user, it may still contain a small amount of novel information (which is, for example, often the case in news reporting). Therefore, they independently evaluate the performance of an information retrieval system with respect to two separate definitions of relevance: a topical relevance and a novel relevance. We believe that this (or a similar) approach is particularly attractive for the field of XML retrieval, where systems tend to retrieve mutually overlapping (and thus redundant) information units. Some aspects of novel relevance are investigated in detail by the TREC Novelty track [18].

## 3.3 Aspect Relevance

A user request often represents a complex information need that may comprise smaller (and possibly independent) parts, often called *aspects*. The goal of an information retrieval system is then to retrieve information units that cover as many aspects of the information need as possible. In this context, *aspect relevance* is defined as topical relevance of the retrieved unit to a particular aspect of the information need, whereas *aspect coverage* is defined as the number of aspects for which relevant retrieved units exist. Zhai [22] describes a formal approach to modelling aspect relevance. INEX uses a somewhat modified definition of aspect relevance, which will be discussed in more detail below.

## 3.4 The INEX Relevance Definition

From 2003 in INEX, the relevance of an information unit (a document or a document component) to a request (a query) is described by two dimensions: *Exhaustivity*, which represents topical relevance that models the extent to which the information unit discusses aspects of the information need represented by the request, and *Specificity*, which also represents topical relevance, but models the extent to which the information unit focuses on aspects of the information need. For example, an information unit may be highly exhaustive to a user request (since it discusses most or all the aspects of the information need), but only marginally specific (since it also focuses on aspects other than those concerning the information need). Conversely, an information unit may be highly specific to a user request (since there is no non-relevant information and it only focuses on aspects concerning the information need), but it may be marginally exhaustive (since it discusses only a few aspects of the information need).

In traditional information retrieval, a binary relevance scale is often used to assess the relevance of an information unit (usually a whole document) to a user request[2]. The rele-

[2]Recent Robust and Web tracks in TREC, however, use a non-binary relevance scale for evaluation.

| Specific | Exhaustive | | | |
|---|---|---|---|---|
| | *Highly* | *Fairly* | *Marginally* | *None* |
| *Highly* | **E3S3** | **E2S3** | **E1S3** | **E0S0** |
| *Fairly* | **E3S2** | **E2S2** | **E1S2** | **E0S0** |
| *Marginally* | **E3S1** | **E2S1** | **E1S1** | **E0S0** |
| *None* | **E0S0** | **E0S0** | **E0S0** | **E0S0** |

**Table 1: The 10-point relevance scale, as adopted by INEX. Each point of the relevance scale combines a particular grade from the Exhaustivity dimension with a corresponding grade from the Specificity dimension.**

vance value of the information unit is restricted to either zero (when the unit is not relevant to the request) or one (when the unit is relevant to the request). INEX, however, adopts a four-graded relevance scale for each of the relevance dimensions, such that the relevance of an information unit to a request ranges from none, to marginally, to fairly, or to highly exhaustive or specific, respectively. To identify *relevant* units of information, that is, units of information that are both exhaustive and specific to a user request, a combination of the grades from each of the two relevance dimensions is used. These relevant units are then, according to INEX, "the most appropriate units of information to return as an answer to the query" [16]. Table 1 shows the combination of the grades from each of the two relevance dimensions, which represents the 10-point relevance scale used by INEX.

The two relevance dimensions, *Exhaustivity* and *Specificity*, are not completely independent. An information unit that is not exhaustive is at the same time not specific to the request (and vice versa), which restricts the space of combining the grades of the two dimensions to ten possible values. In the remainder of the paper, a relevance value of an information unit to a request will be denoted as `EeSs`, where `E` represents *Exhaustivity*, `S` represents *Specificity*, and `e` and `s` represent integer numbers between zero and three. For example, `E1S3` represents an information unit that is marginally exhaustive and highly specific to a request. An information unit is considered relevant only if both `e` and `s` are greater than zero. The relevance value `E0S0` therefore denotes a non-relevant information unit, whereas the value `E3S3` denotes a highly relevant information unit.

### *Comparison with Aspect Relevance*

A strong parallel may be drawn between *Exhaustivity* and *Specificity*, the two INEX relevance dimensions, with *aspect coverage* and *aspect relevance*. *Exhaustivity* maps the aspect coverage to a four-point relevance scale, from `E0` being "the XML element does not discuss the query at all" [16], to `E3` being "the XML element discusses most or all aspects of the query" [16]. *Specificity*, on the other hand, is almost identical to aspect relevance.

## 3.5 Unified Relevance Definition

A notable attempt to construct a unified definition of relevance is given by Mizzaro [13]. He formalises a framework capable of modelling various definitions of relevance by embedding it in a four-dimensional space.

The first dimension deals with the type of entities for which the relevance is defined. It can take one of the following three values: *Document*, *Surrogate*, or *Information*. *Document* refers to the information unit a user will obtain as a result of their search; this may represent a full-text document, an image, video, or, in the case of XML retrieval, a document component. *Surrogate* refers to a form of representation of *Document*; this may be of a set of terms, bibliographic data, or a condensed abstract of the information unit. The third value, *Information*, refers to a rather abstract concept, which depends on the type and amount of information the user receives while reading or consuming the contents of the returned unit of information.

The second dimension relates to the level at which the user request is dealt with. There are four possible levels: *Problem*, *Information need*, *Request*, or *Query*. The *Problem* (also referred to as Real Information Need – RIN [10]) relates to the actual problem that a user is faced with, and for which information is needed to help solve it. The user may not be fully aware of the actual problem; instead, in their minds they perceive it by forming a mental image. This mental image in fact represents the *Information need* (also referred to as Perceived Information Need – PIN [10]). *Request* is a way of communicating the *Information need* to others by specifying it in a natural language. For the *Request* to be recognised by a retrieval system, it needs to be represented by a *Query*. The *Query* usually consists of a set of terms, optionally including phrases or logical query operators.

Relevance can then simply be seen as a combination of any of the entities from the two dimensions above; that is, it can be seen as a combination of any of the *values* from the first dimension with any of the *levels* from the second dimension. Indeed, phrases such as "relevance of a *Surrogate* to a *Query*" or "relevance of a *Document* to a *Request*" are often used. Mizzaro, however, argues that this relevance space does not actually represent the space of all possible relevances. Rather, there is also a third dimension that specifies the nature of the relationship between the two dimensions. The components of this third dimension are *Topic*, *Task*, *Context*, or any combination of the three. The *Topic* (or topical relevance [10]) specifies how similar the two entities are to user's area of interest. For example, if the user is interested in finding information about the overlap problem in XML retrieval, the topical relevance will represent the level of similarity of the retrieved unit to the query with respect to that particular area of interest. The *Task* (or task relevance [10]) specifies the level of usefulness of the information found in an entity for the actual task performed by the user (for example, writing a paper or preparing a lecture). The final component, *Context*, includes everything that is not previously covered by *Topic* and *Task*, but which nevertheless affects the whole process of retrieval (such as search costs, or the amount of novel information found, or anything else).

Since the information seeking process may evolve in time, a fourth dimension, *Time*, is needed to model the fact that users often change their perception of the information they seek to find. For example, at a certain point in time an information unit (*Surrogate* or *Document*) may not be rel-

evant to a user request (*Query* or *Request*), however due to the evolving nature of the seeking process the user may learn something that would permit them to understand the content of the unit, which, in turn, may make the same unit relevant to the request.

A definition of relevance can, therefore, be seen as a point in the above four-dimensional space. Mizzaro [13] argues that the above framework can be used to model and compare different definitions of relevance. For example, the following expression may be used to model the system-oriented definition of relevance described in Section 3.1: *Topical relevance* of a *Surrogate* to a *Query* at a certain point in *Time* (the time when the request was formulated as a query and submitted to the retrieval system). However, finding an expression that may be used to model the INEX definition of relevance turns out to be quite a challenging task. The main problem is that both the INEX relevance dimensions, *Exhaustivity* and *Specificity*, are based on topical relevance, which corresponds to the *Topic* component of the third relevance dimension in the unified framework. We contend that one relevance dimension based on topical relevance should be used, or possibly two orthogonal dimensions that correspond to different components of the above framework. In Section 6 we propose a much simpler definition of relevance, and argue that its corresponding relevance scale would be a better choice for INEX.

# 4. BEHAVIOUR ANALYSIS FOR BACKGROUND TOPICS

In this section, we separately analyse the assessor's and users' behaviour when judging the relevance of returned elements for the *Background* topic B1. In order to identify the best retrieval elements for this topic, we also analyse and compare the level of agreement between the assessor and the users.

## 4.1 Analysis of Assessor's Behaviour

Figure 3 shows an analysis of the relevance judgements for topic B1 (the INEX 2004 CO topic 192) that were obtained from one assessor. As shown in the figure, we use only four element names in our analysis: `article`, `sec`, `ss1`, and `ss2`. The $x$-axis contains the 9-point relevance scale which is a result of combining the grades of the two INEX relevance dimensions (the case `E0S0` is not shown). The $y$-axis contains the number of occurrences of relevant elements for each point of the relevance scale. For a relevance point, the number of occurrences of each of the four element names is also shown.

The total number of relevant elements for topic B1 is 32. Of these, 11 elements have been judged as `E2S1`, nine as `E1S1`, six as `E2S3`, two as `E3S3` or `E2S2`, and one as `E3S1` or `E1S2`. Interestingly, none of the relevant elements have been judged as either `E3S2` or `E1S3`. The number of occurrences of the four element names is as follows. The `sec` elements occur most frequently with 18 occurrences, followed by `article` with ten, `ss1` with three, and `ss2` with one occurrence, respectively. The total number of elements that have been judged as non-relevant (`E0S0`) for topic B1 is 1158, of which 513 are `sec` elements, 411 are `ss1`, 186 are `article`, and 48 are `ss2` elements.

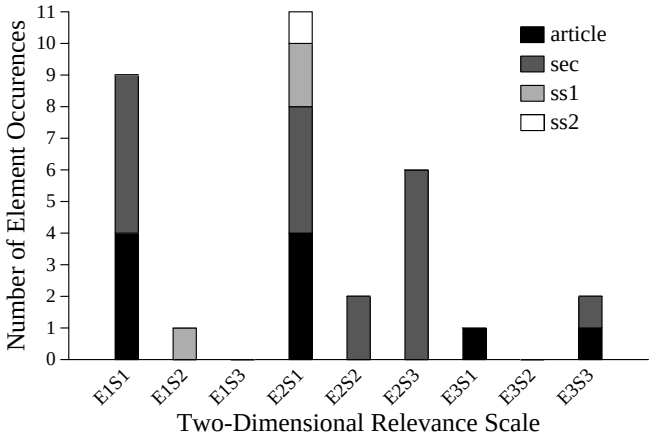


**Figure 3: Analysis of assessor's behaviour for topic B1. For each point of the relevance scale, the figure shows the total number of relevant elements, and the number of relevant elements for each of the element names.**

### *Level of Overlap*

The above statistics show that the `E2S1` and `E1S1` points of the relevance scale contain around 63% of the relevant elements for topic B1. Moreover, further analysis reveals that there is a substantial amount of overlap among these elements. More precisely, there is 64% set-based overlap among the 11 `E2S1` elements, where the four `article` elements contain all of the section and sub-section elements. Similarly, there is 56% overlap among `E1S1` elements, where of nine elements, four `article` elements contain the other five `sec` elements. Interestingly, the other points of the relevance scale do not suffer from overlap. The two highly relevant elements (`E3S3`), for example, belong to different XML files.

### *Correlation between Relevance Grades*

In the following we investigate the correlation between the grades of the two relevance dimensions for topic B1. We want to check whether, while judging relevant elements, the assessor's choice of combining the grades of the two relevance dimensions is influenced by a common aspect [9].

The top half of Table 2 shows the correlation between the grades of the two relevance dimensions for topic B1, as judged by the assessor. For each grade of the *Exhaustivity* relevance dimension (columns), the value of `Sp|Ex` shows the percentage of the cases where an element is judged as `Sp` (specific), given that it has already been judged as `Ex` (exhaustive). Similarly, for each grade of the *Specificity* relevance dimension (rows), the value of `Ex|Sp` shows the percentage of the cases where an element is judged as `Ex` (exhaustive), given that it has already been judged as `Sp` (specific). For example, the `Sp|Ex` value of column `E3` and row `S3` is 66.67, indicating that in 66.67% of the cases a highly exhaustive element is also judged as highly specific. We now analyse the correlation between the grades of each separate relevance dimension.

For *Exhaustivity*, we observe that in 90% of the cases a marginally exhaustive (`E1`) element is also judged as margi-

| Assessor: | Specificity | Exhaustivity E3 Sp\|Ex (%) | E3 Ex\|Sp (%) | E2 Sp\|Ex (%) | E2 Ex\|Sp (%) | E1 Sp\|Ex (%) | E1 Ex\|Sp (%) |
|---|---|---|---|---|---|---|---|
| | *S3* | **66.67** | 25.00 | 31.57 | *75.00* | 0.00 | 0.00 |
| | *S2* | 0.00 | 0.00 | 10.53 | *66.67* | 10.00 | 33.33 |
| | *S1* | 33.33 | 4.62 | **57.90** | *52.38* | **90.00** | 43.00 |

| Users: | Specificity | Exhaustivity E3 Sp\|Ex (%) | E3 Ex\|Sp (%) | E2 Sp\|Ex (%) | E2 Ex\|Sp (%) | E1 Sp\|Ex (%) | E1 Ex\|Sp (%) |
|---|---|---|---|---|---|---|---|
| | *S3* | **69.62** | *69.62* | 36.84 | 22.15 | 12.26 | 8.23 |
| | *S2* | 27.22 | *41.34* | **40.00** | 36.54 | 21.70 | 22.12 |
| | *S1* | 3.16 | 5.16 | 23.16 | 22.68 | **66.04** | *72.16* |

**Table 2: Correlation between the grades of the two relevance dimensions for topic B1, as judged by both the assessor and the users. Depending on the relevance dimension, the highest correlation of each grade is shown either in bold (for Exhaustivity) or italics (for Specificity).**

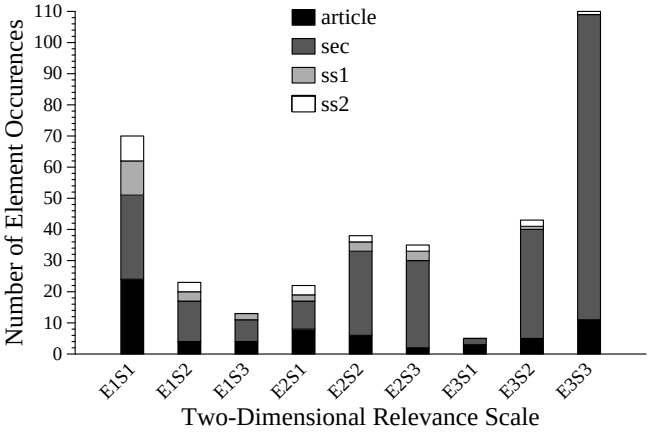


**Figure 4: Analysis of users' behaviour for topic B1. For each point of the relevance scale, the figure shows the total number of relevant elements, and the number of relevant elements for each of the element names.**

nally specific (*S1*). This is somehow intuitive, since by definition a marginally exhaustive element discusses only a few aspects of the information need, so its focus may be on aspects other than those concerning the information need. However, for topic B1, the number of `E1` elements is around 30% of the total number of relevant elements, so the above correlation should be treated carefully. In contrast, the number of fairly exhaustive elements (`E2`) is around 60% of the total number of relevant elements, and in 58% of the cases a fairly exhaustive element is (again) judged as *S1*. For highly exhaustive (`E3`) elements, we find that in 67% of the cases an `E3` element is also judged as highly specific (*S3*), although the number of `E3` elements is very low (only 10% of the total number of relevant elements).

For *Specificity*, the number of marginally specific (*S1*) elements is around 66% of the total number of relevant elements, where in 52% of the cases an *S1* element is judged as fairly exhaustive (`E2`), while in 43% of the cases it is judged as marginally exhaustive (`E1`). Fairly specific (*S2*) elements are 9% of the total number of relevant elements, and in 67% of the cases an *S2* element is judged as `E2`. Finally, in 75% of the cases a highly specific (*S3*) element is (again) judged as `E2`, although the number of highly specific elements is around 25% of the total number of relevant elements.

## 4.2 Analysis of Users' Behaviour

Figure 4 shows the relevance judgements for topic B1 that were obtained from 50 users. Unlike in the assessor's case, an element may have been judged by more than one user, so each relevance point in Figure 4 may contain multiple occurrences of a given element.

The total number of occurrences of relevant elements for topic B1 is 359. Around 61% of this number are elements that have been judged either as `E3S3` (110), `E1S1` (70), or `E2S2` (38). All the 10 points of the relevance scale were used by users. However, different number of users have judged elements for each relevance point. For example, 41 (out of 50) users have judged at least one element as `E3S3`, whereas this number is 35 for `E1S1`, 23 for `E2S2`, and 20 and below for the other points of the relevance scale. The `sec` elements occur most frequently with 246 occurrences, followed by `article` with 67, `ss1` with 25, and `ss2` with 21 occurrences, respectively. The total number of element occurrences judged as non-relevant (`E0S0`) for topic B1 is 181, of which 80 are `sec` elements, 72 are `article`, 26 are `ss1`, and only 3 are `ss2` elements. Also, 39 (out of 50) users have judged at least one element as `E0S0`.

### *Level of Overlap*

A more detailed analysis of the user judgements for topic B1 reveals that there is almost no overlap among the elements that belong to any of the nine points of the relevance scale. More precisely, there is 14% set-based overlap among the 110 `E3S3` elements, 0% overlap among the 70 `E1S1` elements, and 0% overlap for the other seven points of the relevance scale. The above finding therefore confirms the hypothesis that users do not want to retrieve (and thus do not tolerate) redundant information.

### *Correlation between Relevance Grades*

The lower half of Table 2 shows the correlation between the grades of the two relevance dimensions for topic B1, as judged by users. For both *Exhaustivity* and *Specificity*, two strong correlations are visible. First, in 66% of the cases a marginally exhaustive (`E1`) element is also judged as marginally specific (*S1*) (and vice versa). Second, in 70%

| Assessor | | User judgements | | | | | | | | | | | Agreement |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Judgement | Total | E3S3 | E3S2 | E3S1 | E2S3 | E2S2 | E2S1 | E1S3 | E1S2 | E1S1 | E0S0 | Total | (%) |
| E3S3 | 2 | 25 | 10 | 0 | 5 | 4 | 1 | 0 | 2 | 1 | 0 | 48 (2) | 52.08 |
| E3S2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 (0) | 0.00 |
| E3S1 | 1 | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 3 (1) | 0.00 |
| E2S3 | 6 | 60 | 14 | 1 | 18 | 13 | 4 | 3 | 7 | 8 | 0 | 128 (6) | 14.06 |
| E2S2 | 2 | 14 | 4 | 1 | 2 | 1 | 0 | 1 | 0 | 0 | 1 | 24 (1) | 4.17 |
| E2S1 | 11 | 1 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 2 | 0 | 6 (3) | 0.00 |
| E1S3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 (0) | 0.00 |
| E1S2 | 1 | 0 | 0 | 0 | 2 | 1 | 0 | 0 | 0 | 1 | 0 | 4 (1) | 0.00 |
| E1S1 | 9 | 3 | 2 | 1 | 1 | 7 | 3 | 0 | 2 | 11 | 17 | 47 (5) | 23.40 |
| E0S0 | 1158 | 1 | 6 | 2 | 2 | 7 | 9 | 7 | 6 | 36 | 99 | 175 (59) | 56.57 |
| Total | 1190 | 105 | 36 | 5 | 32 | 34 | 19 | 11 | 17 | 59 | 117 | 435 (78) | **15.10** |

**Table 3: The level of agreement between the assessor and the users for topic B1. For each point of the relevance scale, the percentage of users that agree with the assessor's judgements of corresponding elements is shown. Numbers in brackets represent numbers of unique elements judged by users. The overall level of agreement for topic B1 is shown in bold.**

of the cases a highly exhaustive (E3) element is also judged as highly specific (*S3*) (and vice versa). The number of E1 elements is around 30% of the total number of relevant elements, whereas 44% of the total number of relevant elements are E3 elements. The number of *S1* and *S3* elements is almost the same as the number of E1 and E3 elements, respectively. No strong correlations are, however, visible in the case of E2 and *S2* elements.

## 4.3 Analysis of the Level of Agreement

The analysis of the level of agreement concerns the amount of information identified as relevant by *both* the assessor and the users. The aim of this analysis is to identify the best units of retrieval for topic B1.

Table 3 shows the level of agreement between the assessor and the users for each point of the relevance scale. The two columns on the left refer to the assessor's judgements, where for each relevance point (the Judgement column), the total number of judged elements that belong to this point is shown (the Total column). The values in the User Judgements columns show how users actually judged any (or all) of the corresponding elements judged by the assessor. The Total column on the right shows the total number of user judgements for each point of the relevance scale. Numbers in brackets represent numbers of unique elements judged by users. The Agreement column shows the level of agreement between the assessor and the users, where the percentage is calculated for each relevance point.

For example, the first row in the table indicates that there are two elements judged as E3S3 by the assessor, and that of 48 total user judgements, there are 25 cases when users judged any (or both) of these two elements as E3S3, ten cases as E3S2, five cases as E2S3, and so on. The level of agreement between the assessor and the users for the E3S3 point of the relevance scale is 52.08% (since in 25 out of 48 cases users judged these elements as E3S3). Note that for this relevance point we only consider the user judgements made on two unique elements, which correspond to the same elements judged as E3S3 by the assessor. As shown in the table, the overall level of agreement between the assessor and the users for topic B1 is 15%.

Several observations can be made from the statistics shown in Table 3.

First, users judged 19 (unique) of the 32 *relevant* elements as identified by the assessor for topic B1. In 7% of the cases, however, users judged some of these elements to be *not relevant*. Conversely, 59 (unique) of the 1,158 *non-relevant* elements, as identified by the assessor, were also judged by users, and in 43% of the cases users judged some of those elements to be *relevant*.

Second, the highest level of agreement between the assessor and the users is on highly relevant (E3S3) and non-relevant (E0S0) elements, with agreement values of 52% and 57%, respectively. This shows that both the assessor and the users clearly perceive the end points of the relevance scale. However, the other points of the relevance scale are not perceived as well. For example, although the highest number of user judgements is on the E2S3 relevance point (around 50%), in only 14% of the cases users actually judged these elements as E2S3. In fact, in the majority of the cases (47%), the users judged these elements to be highly relevant (E3S3). Similar observations can be made for the E1S1 relevance point, where in 36% of the cases the users judged these elements to be non-relevant (E0S0). Note that, even though the number of judged E3S3 and E1S1 elements is roughly the same, the level of agreement for the E3S3 relevance point is more than two times greater than the level of agreement for the E1S1 relevance point.

Last, a more detailed analysis of the above statistics reveals that the agreement between the assessor and the users is almost the same for each separate relevance dimension. More precisely, the overall agreement for *Exhaustivity* is 45%, whereas the overall agreement for *Specificity* is 44%. The agreement for highly exhaustive (E3) elements is 71%, where 20% of the total number of confirmed *relevant* elements is on E3 elements. On the other hand, the agreement for highly specific (S3) elements is 63%, where 68% of the confirmed relevant elements are S3 elements. This shows that although the number of user judgements for the S3 grade is more than three times greater than the number of judgements for the E3 grade, highly exhaustive elements are perceived better than highly specific elements.

**File: cg/1998/g1016**

| Assessor | | User judgements | | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Element | Judgement | E3S3 | E3S2 | E3S1 | E2S3 | E2S2 | E2S1 | E1S3 | E1S2 | E1S1 | E0S0 | (users) |
| /article[1] | E3S3 | 9 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 12 |
| //bdy[1]/sec[2] | E2S3 | 9 | 5 | 1 | 7 | 6 | 2 | 1 | 2 | 2 | 0 | 35 |
| //bdy[1]/sec[3] | E2S2 | 14 | 4 | 1 | 2 | 1 | 0 | 1 | 0 | 0 | 1 | 24 |
| //bdy[1]/sec[4] | E2S3 | 19 | 0 | 0 | 4 | 1 | 1 | 0 | 0 | 2 | 0 | 27 |
| //bdy[1]/sec[5] | E2S3 | 18 | 3 | 0 | 3 | 2 | 1 | 0 | 2 | 1 | 0 | 30 |
| //bdy[1]/sec[6] | E2S3 | 8 | 2 | 0 | 2 | 1 | 0 | 1 | 0 | 1 | 0 | 15 |
| //bdy[1]/sec[7] | E2S3 | 6 | 4 | 0 | 2 | 2 | 0 | 1 | 3 | 2 | 0 | 20 |

**File: cg/1995/g5095**

| Assessor | | User judgements | | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Element | Judgement | E3S3 | E3S2 | E3S1 | E2S3 | E2S2 | E2S1 | E1S3 | E1S2 | E1S1 | E0S0 | (users) |
| /article[1] | E3S1 | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 3 |
| //bdy[1]/sec[1] | E3S3 | 16 | 7 | 0 | 5 | 4 | 1 | 0 | 2 | 1 | 0 | 36 |
| //bdy[1]/sec[2] | E0S0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 3 |
| //bdy[1]/sec[3] | E0S0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 2 |
| //bdy[1]/sec[4] | E0S0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 3 |

**Table 4: Distribution of relevance judgements for the XML files *cg/1998/g1016* (top) and *cg/1995/g5095* (bottom) for topic B1. For each element, the assessor judgement and the distribution of users' judgements are shown. The total number of users who judged a particular element is listed in the last column.**

### *Best Units of Retrieval*

Previous analysis shows that of all the *relevant* elements as judged by users, the E3S3 point of the relevance scale has the highest level of agreement. There are two elements judged as highly relevant by the assessor for topic B1 – one article and one sec – that belong to different XML files. The article element belongs to file cg/1998/g1016, while the sec element belongs to cg/1995/g5095. We are interested in finding in these files the *best units of retrieval* for topic B1. In the following analysis, we examine the retrieval behaviour of both the assessor and the users for each of these files.

Table 4 shows the distribution of relevance judgements for relevant elements in the two XML files, as done by both the assessor and the users. The two columns on the left refer to the assessor, where for each relevant element in the file (the Element column), the assessor's judgement is also shown (the Judgement column). The values in the User Judgements columns show the distribution of users' judgements for each particular element; that is, the number below each relevance point represents the number of users that judged that element. The total number of users who judged a particular element is shown in the Total column.

For the file cg/1998/g1016, the top half of the table shows that the highly relevant (E3S3) article element was judged by 12 (out of 50) users, and that 75% of them confirmed it to also be highly relevant. Interestingly, around 70% of the relevant elements in this file have been judged as E2S3 by the assessor, and there were 25 users (on average) who have also judged these elements. However, there is only a 14% agreement (on average) between the assessor and the users for the E2S3 relevance point. In fact, if we take a closer look at the user judgements, we see that most users judged the E2S3 elements to be highly relevant (E3S3) elements. For example, there were 27 users in total who judged the sec[4] element (judged as E2S3 by the assessor), and 70% of them judged this element to be highly relevant (E3S3).

The above analysis shows that the agreement between the users and the assessor on the *best units of retrieval* for the file cg/1998/g1016 is not exact. Further analysis confirms that the level of agreement between the assessor and the users is greater for highly exhaustive elements than for highly specific ones. More precisely, although the number of user judgements for the S3 grade is more than ten times greater than the number of judgements for the E3 grade, there is a 65% agreement for highly specific elements, while there is a 100% agreement for highly exhaustive elements.

For the file cg/1995/g5095, the lower half of Table 4 shows that there are only two elements identified as *relevant* by the assessor, which makes it impossible to draw any sound conclusions. The highly relevant sec element was judged by 36 (out of 50) users, and around 45% of the users also confirmed it to be highly relevant. Interestingly, three sec elements were judged as not relevant by the assessor, and almost all of the users who judged these elements also confirm them to be non-relevant.

## 5. BEHAVIOUR ANALYSIS FOR COMPARISON TOPICS

In this section, we separately analyse the assessor's and users' behaviour when judging the relevance of returned elements for the *Comparison* topic C2. In order to identify the best retrieval elements for this topic, we also analyse and compare the level of agreement between the assessor and the users.

### 5.1 Analysis of Assessor's Behaviour

Figure 5 shows the relevance judgements for the INEX 2004 CO topic 198 (topic C2) that were obtained from one assessor. As shown in the figure, the total number of relevant elements for topic C2 is 153, of which the majority (81%)

| Assessor: | Specificity | Exhaustivity E3 Sp\|Ex (%) | E3 Ex\|Sp (%) | E2 Sp\|Ex (%) | E2 Ex\|Sp (%) | E1 Sp\|Ex (%) | E1 Ex\|Sp (%) |
|---|---|---|---|---|---|---|---|
| | *S3* | **100.00** | 33.33 | 22.22 | 33.33 | 1.41 | 33.33 |
| | *S2* | 0.00 | 0.00 | **66.67** | 27.27 | 11.27 | *72.73* |
| | *S1* | 0.00 | 0.00 | 11.11 | 0.80 | **87.32** | *99.20* |

| Users: | Specificity | Exhaustivity E3 Sp\|Ex (%) | E3 Ex\|Sp (%) | E2 Sp\|Ex (%) | E2 Ex\|Sp (%) | E1 Sp\|Ex (%) | E1 Ex\|Sp (%) |
|---|---|---|---|---|---|---|---|
| | *S3* | **52.99** | *48.84* | 30.13 | 32.56 | 13.77 | 18.60 |
| | *S2* | 35.04 | 27.33 | **43.59** | *43.33* | 27.54 | 29.33 |
| | *S1* | 11.97 | 8.50 | 26.28 | 27.45 | **58.68** | *64.05* |

**Table 5: Correlation between the grades of the two relevance dimensions for topic C2, as judged by both the assessor and the users. Depending on the relevance dimension, the highest correlation of each grade is shown either in bold (for Exhaustivity) or italics (for Specificity).**

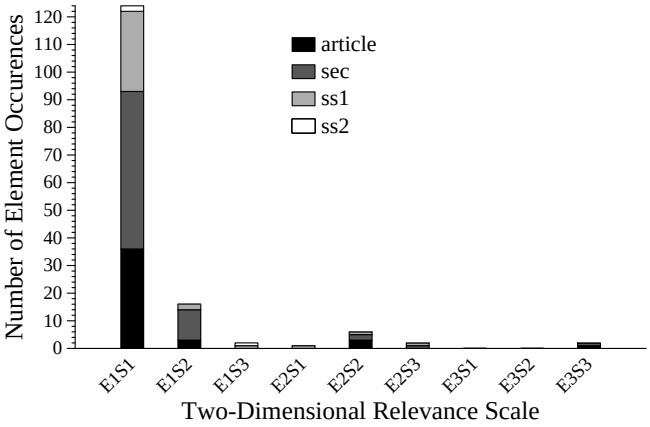


**Figure 5: Analysis of assessor's behaviour for topic C2. For each point of the relevance scale, the figure shows the total number of relevant elements, and the number of relevant elements for each of the element names.**

have been judged as E1S1. Interestingly, none of the relevant elements have been judged as either E3S2 or E3S1. The distribution of the four element names is as follows. The sec elements occur most frequently with 72 occurrences, followed by article with 43, ss1 with 35, and ss2 with only three occurrences, respectively. The total number of elements that have been judged as non-relevant (E0S0) for topic C2 is 1094, of which 547 are sec elements, 304 are ss1, 191 are article, and 52 are ss2 elements.

### *Level of Overlap*

The above statistics show that the E1S1 point of the relevance scale contains almost all of the relevant elements for topic C2. However, as for the topic B1, there is a substantial overlap among these elements. More precisely, there is a 63% set-based overlap among the 124 E1S1 elements. On the other hand, the other points of the relevance scale – except the E3S3 point – do not suffer from overlap. For the E3S3 point, there is a 50% set-based overlap, where the two highly relevant elements (one article and one sec) belong to the same XML file.

### *Correlation between Relevance Grades*

The top half of Table 5 shows the correlation between the grades of the two relevance dimensions for topic C2, as judged by the assessor. We observe that each of the three grades of the *Exhaustivity* dimension is strongly correlated with its corresponding grade of the *Specificity* dimension. This is most evident for the E1 grade, where in 87% of the cases a marginally exhaustive (E1) element is also judged as marginally specific (*S1*). The number of E1 elements is 93% of the total number of relevant elements. The same is not true for the grades of the Specificity dimension, however, where both the S2 and S1 grades are strongly correlated with the E1 grade. Most notably, in 99% of the cases a marginally specific (S1) element is also judged as marginally exhaustive (*E1*), where the number of S1 elements is 82% of the total number of relevant elements.

## 5.2 Analysis of Users' Behaviour

Figure 6 shows the relevance judgements for topic C2 that were obtained from 52 users. As shown in the figure, the total number of occurrences of relevant elements is 445, of which around half of that number are elements that belong to the following three points of the relevance scale: E1S1 (101), E2S2 (66), and E3S3 (63). Interestingly, approximately the same number of users (34 out of 52) judged at least one element that belongs to each of these three points. In contrast, 22 users (on average) judged at least one element that belongs to the other six points of the relevance scale.

The distribution of the four element names is as follows. The sec and article elements occur most frequently with 159 and 153 occurrences, followed by ss1 elements with 130, and ss2 elements with only three occurrences, respectively. The total number of element occurrences judged as non-relevant (E0S0) for topic C2 is 170, of which 116 are sec elements, 27 are ss1, 26 are article, and only one element is an ss2 element. Also, 38 out of 52 users have judged at least one element as E0S0.

### *Level of Overlap*

Further analysis of the user judgements for topic C2 reveals that there is almost no overlap among the elements that belong to any of the nine points of the relevance scale. More specifically, there is 3% set-based overlap for the E1S1 point,

| Assessor | | Users | | | | | | | | | | | Agreement |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Relevance | Total | E3S3 | E3S2 | E3S1 | E2S3 | E2S2 | E2S1 | E1S3 | E1S2 | E1S1 | E0S0 | Total | (%) |
| E3S3 | 2 | 6 | 4 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 14 (2) | 42.86 |
| E3S2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 (0) | 0.00 |
| E3S1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 (0) | 0.00 |
| E2S3 | 2 | 16 | 4 | 1 | 6 | 7 | 1 | 0 | 0 | 2 | 1 | 38 (2) | 15.79 |
| E2S2 | 6 | 20 | 8 | 0 | 12 | 12 | 6 | 3 | 5 | 7 | 3 | 76 (5) | 15.79 |
| E2S1 | 1 | 2 | 0 | 1 | 1 | 4 | 0 | 0 | 1 | 1 | 0 | 10 (1) | 0.00 |
| E1S3 | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 4 (2) | 25.00 |
| E1S2 | 16 | 1 | 1 | 1 | 2 | 1 | 0 | 1 | 2 | 1 | 2 | 12 (7) | 16.67 |
| E1S1 | 124 | 17 | 19 | 6 | 16 | 24 | 16 | 8 | 18 | 45 | 38 | 207 (34) | 21.74 |
| E0S0 | 1094 | 2 | 2 | 2 | 3 | 3 | 10 | 5 | 9 | 25 | 85 | 146 (52) | 58.22 |
| Total | 1247 | 64 | 38 | 12 | 41 | 52 | 33 | 19 | 35 | 83 | 130 | 507 (105) | **19.61** |

**Table 6: The level of agreement between the assessor and the users for topic C2. For each point of the relevance scale, the percentage of users that agree with the assessor's judgements of corresponding elements is shown. Numbers in brackets represent numbers of unique elements judged by users. The overall level of agreement for topic C2 is shown in bold.**

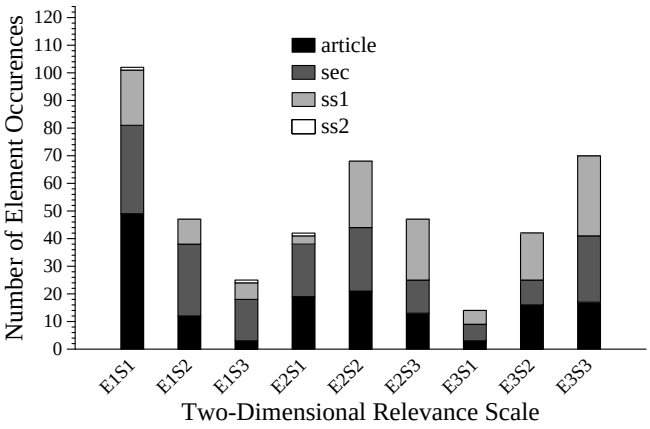


**Figure 6: Analysis of users' behaviour for topic C2. For each point of the relevance scale, the figure shows the total number of relevant elements, and the number of relevant elements for each of the element names.**

0% for the E2S2 point, 9% overlap for the E3S3 point, and 0% overlap for the other six points of the relevance scale.

### *Correlation between Relevance Grades*

The lower half of Table 5 shows the correlation between the grades of the two relevance dimensions for topic C2, as judged by users. Although no strong correlations are visible, the values in the table show that, as in assessor's case, the highest correlations are between the same grades of each of the two relevance dimensions.

## 5.3 Analysis of the Level of Agreement

In this section we analyse the amount of information identified as relevant by *both* the assessor and the users. Table 6 shows the level of agreement between the assessor and the users for each point of the relevance scale. Three observations can be made from the statistics shown in the table.

First, users judged 53 (unique) of the 153 *relevant* elements as identified by the assessor for topic C2. In 12% of the cases, however, users judged these elements to be *not relevant*. Conversely, 52 (unique) of the 1094 *non-relevant* elements, as identified by the assessor, were also judged by users, and in 42% of the cases users judged these elements to be *relevant*.

Second, as for topic B1 the highest level of agreement between the assessor and the users is on the end points of the relevance scale: E3S3 (43%) and E0S0 (58%), although the number of user judgements for the E3S3 relevance point is much less than the number of judgements for the E0S0 point. The E1S1 relevance point has the highest number of user judgements (207 out of 507), and in 22% of the cases users also judged these elements to be E1S1. Also, there are 76 user judgements for the E2S2 relevance point, however in 26% of the cases users actually judged the E2S2 elements to be highly relevant (E3S3) elements.

Third, a more detailed analysis shows that the level of agreement between the assessor and the users differs for each separate relevance dimension. More precisely, the overall agreement for *Exhaustivity* is 53%, while the overall agreement for *Specificity* is 45%. The agreement for highly exhaustive (E3) elements is 79%, and 4% of the total number of confirmed *relevant* elements is on E3 elements. In contrast, the agreement for highly specific (S3) elements is 55%, where 18% of confirmed relevant elements are S3 elements. This shows that, as for topic B1, highly exhaustive elements are perceived better than highly specific elements.

### *Best Units of Retrieval*

There are two elements judged as highly relevant by the assessor for topic C2, one article and one sec, which belong to the same XML file: co/2000/rx023. To identify the best units of retrieval, in the following we examine the behaviour of both the assessor and the users for this file.

Table 7 shows the distribution of relevance judgements for relevant elements in the XML file co/2000/rx023, as done by both the assessor and the users. As shown in the table, the two highly relevant (E3S3) elements were judged by the same number of users (seven out of 52). Of the users that judged each of these elements, 57% confirmed the article[1] to be highly relevant, while only 29% confirmed the sec[3] element to be highly relevant. Many users, however, found

**File: co/2000/rx023**

| Assessor | | User judgements | | | | | | | | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Element | Judgement | E3S3 | E3S2 | E3S1 | E2S3 | E2S2 | E2S1 | E1S3 | E1S2 | E1S1 | E0S0 | (users) |
| /article[1] | E3S3 | 4 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 7 |
| //bdy[1]/sec[1] | E2S2 | 5 | 0 | 0 | 1 | 2 | 2 | 1 | 3 | 1 | 2 | 17 |
| //bdy[1]/sec[2] | E2S2 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 5 |
| //bdy[1]/sec[3] | E3S3 | 2 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 7 |
| //bdy[1]/sec[3]/ss1[1] | E2S2 | 10 | 3 | 0 | 8 | 6 | 2 | 2 | 1 | 1 | 0 | 33 |
| //bdy[1]/sec[3]/ss1[2] | E2S1 | 2 | 0 | 1 | 1 | 4 | 0 | 0 | 1 | 1 | 0 | 10 |
| //bdy[1]/sec[3]/ss1[4] | E1S1 | 3 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 7 |
| //bdy[1]/sec[3]/ss1[5] | E2S3 | 7 | 3 | 1 | 4 | 2 | 0 | 0 | 0 | 0 | 0 | 17 |
| //bdy[1]/sec[3]/ss1[6] | E1S3 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 2 |
| //bdy[1]/sec[4] | E2S3 | 9 | 1 | 0 | 2 | 5 | 1 | 0 | 0 | 2 | 1 | 21 |
| //bm[1]/app[1]/sec[1] | E1S1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 3 | 5 |

**Table 7: Distribution of relevance judgements for the XML file *co/2000/rx023* for topic C2. For each element, the assessor judgement and the distribution of users' judgements are shown. The total number of users who judged a particular element is listed in the last column.**

the child elements of the `sec[3]` element (such as `ss1[1]`, `ss1[4]` and `ss1[5]`) to be highly relevant.

From the above distribution of relevance judgements it is hard to draw any sound conclusions as to which elements constitute *best units of retrieval* for this file. Further analysis of the two behaviours for this file again confirms that, for topic C2, the level of agreement between the assessor and the users is greater for highly exhaustive than for highly specific elements. Specifically, although the number of user judgements for the `S3` grade is four times greater than the number of judgements for the `E3` grade, the agreement for highly specific elements is 56%, while there is a 79% agreement for highly exhaustive elements.

## 6. DISCUSSION

In previous sections we separately studied the behaviour of the assessor and the users when judging the relevance of returned elements. We also analysed the level of agreement between the assessor and the users in order to identify the best units of retrieval for each of the two topics.

According to the assessor, most of the relevant elements for topic B1 reside in the `E2S1` and `E1S1` points of the relevance scale. The `E1S1` relevance point also contains most of the relevant elements for topic C2. In both topic cases, however, there is a substantial overlap among these relevant elements: 60% for topic B1, and 63% for topic C2. There are no visible correlations between the grades of each relevance dimension for the assessor of topic B1, whereas for for the assessor of topic C2 each of the three grades of the *Exhaustivity* dimension is strongly correlated with its corresponding grade of the *Specificity* dimension.

According to users, most of the relevant elements in both topic cases reside in the `E1S1`, `E2S2`, and `E3S3` relevance points. Moreover, there is almost no overlap among the relevant elements. Unlike in the assessor's case, the highest correlations between the grades of the relevance dimensions are between the same grades of each of the two dimensions, irrespective of the choice of the topic used. This shows that the two INEX relevance dimensions are not perceived as orthogonal dimensions; in fact, users behave as if each of the grades from either dimension belongs to only one relevance dimension.

The latter finding suggests that the *common aspect* influencing the choice of combining grades from the two INEX relevance dimensions is the fact that the users can not make a clear distinction between the two dimensions (since they are both based on topical relevance). However, it does not mean that the two INEX relevance dimensions are the same. On the contrary, from the *Exhaustivity* definition, higher aspect coverage does not imply that there is less non-relevant information in an element, which means there is no one-to-one correspondence between the two INEX dimensions. Rather, the users' perception – which was empirically identified in this study – suggests that the cognitive load of simultaneously choosing the grades for *Exhaustivity* and *Specificity* is too difficult a task. Part of the problem may be that the users (and the assessor) may not have understood an important property of the *Specificity* dimension: an element should be judged as *highly specific (S3)* if it *does not* contain *non-relevant* information.

The low level of overlap between the judged elements in the users' case shows that *retrieving overlapping units of information is not what users really want.* However, the higher level of overlap in the assessor's case does not necessarily mean that the assessor's behaviour is very different from that of users; indeed, there are *at least* two external factors that may have influenced the observed level of overlap for the assessor:

- The assessor was required to judge many more elements than the users, in order for the obtained relevance judgements to be as exhaustive (and as consistent) as possible; and
- The assessor and the users used different system interfaces, which may have introduced a bias in the way the elements were judged.

The highest level of agreement between the assessor and the users in both topic cases is respectively on highly relevant (`E3S3`) and non-relevant (`E0S0`) elements, which shows

that both the assessor and the users clearly perceive the end points of the relevance scale. However, *the other points of 10-point relevance scale were not perceived as well*. When the two relevance dimensions were analysed separately, we observed that – in both topic cases – *Exhaustivity* is perceived better than *Specificity*.

The above findings suggest that a much simpler relevance scale, and therefore, a much simpler relevance definition, would be a preferable choice for INEX. In the following we propose one such definition of relevance.

## Aspects and Dimensions of Relevance

There are three aspects on which our new definition of relevance is based on:

- There should be only *one* dimension of relevance based on *topical relevance* (rather than two);
- The relevance dimension should use a *binary* relevance scale (rather than graded relevance scale), which determines whether a unit of information is *relevant* or *not* to an information need; and
- There should be second *orthogonal* dimension of relevance, based on the hierarchical relationships among the units of information in XML documents.

The first aspect makes the new relevance definition much simpler than the current one, and more importantly, enables a straightforward integration in the unified relevance framework [13]. The second aspect is directly inspired by the analysis of the level of agreement between the assessor and the users; indeed, the highest level of agreement was shown to be either on *highly relevant* or on *non-relevant* units of retrieval. This means that both the assessor and users clearly agree upon the *binary nature* of topical relevance of the retrieved units, indicating that a unit is either *relevant* or *not* to an information need. The second dimension of relevance, as introduced in the third aspect above, is completely orthogonal to the first dimension. It is defined as follows.

The extent to which a unit of information is relevant to an information need is measured by considering the *difference* between:

- The extent to which aspects of the information need are covered within the unit; and
- The extent to which these aspects are covered within the other *related* units (ancestors or descendants) in the document hierarchy.

For example, a relevant information unit is *just right* to an information need if it mainly just covers aspects of the information need. Alternatively, the information unit can be either *too broad* or *too narrow* to the information need. A relevant information unit is *too broad* if there is a *descendant* that mainly just covers aspects of the information need. Conversely, a relevant information unit is *too narrow* if there is an *ascendant* that is *just right*.

The second dimension of relevance, as defined above, is very similar to *document coverage* used in INEX 2002 [9]. Indeed, document (or component) coverage was used as a relevance dimension in INEX 2002 to measure how specific (or focused) the unit of retrieval is to the information need. List and de Vries [11] describe a formal approach to modelling the document coverage. Similar to our second dimension, some aspects of document coverage depend on the context where the information unit resides, stating that "the component is too small to act as a meaningful unit of information when retrieved by itself" [9]. This, however, makes the document coverage to also be dependent on the size of the retrieved unit. The size of the unit of retrieval, on the other hand, is not explicitly considered in our relevance dimension.

## New Relevance Definition for XML Retrieval

Considering the above observations, we propose the following definition of relevance:

- An information unit is *not relevant* to an information need if it does not cover any of the aspects of the information need;
- An information unit is *relevant* to an information need if it covers any of the aspects of the information need. The extent to which the unit is relevant to the information need can be one of the following:
  - *Broad*, if the unit is too broad and includes other, non-relevant information;
  - *Narrow*, if the unit is too narrow and is part of a larger unit that better covers aspects of the information need; and
  - *Just right*, if the unit mainly just covers aspects of the information need.

The above relevance definition has the following properties:

- In any one document path from the root element to a leaf, *at most* one element can be *Just right*. However, multiple *Just right* elements can exist in an XML document if they belong to different paths;
- Every element in a path that resides *above* the *Just right* element is too broad, and only such elements are considered to be too broad; and
- Every element considered to be too narrow is either a *child* of an element that is *Just right*, or a child of an element that is too narrow. Also, not every child of a relevant element has to be relevant.

There are two relevance dimensions described by the above definition: one based on the topical relevance, which uses a binary relevance scale (*relevant* or *non-relevant*); and another based on hierarchical relationships among the information units in XML documents, which uses a three-graded relevance scale (*Broad*, *Narrow*, or *Just right*).

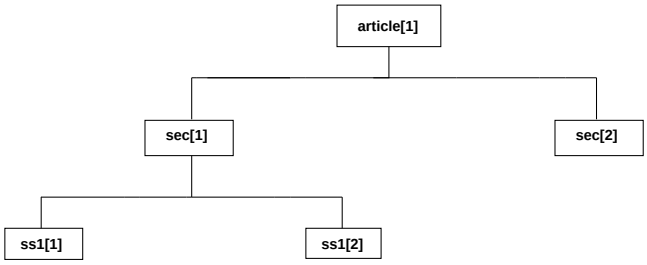


**Figure 7: A representation of an XML document**

### *Example Scenarios*

We further explain the new relevance definition with several example scenarios, with reference to the XML document representation in Figure 7.

*Scenario 1*: Assume that only `ss1[1]` is relevant to an information need, and that it mainly just covers aspects of the information need. Because of the hierarchical relationships between the elements in the above document, both `sec[1]` and `article[1]` will also be relevant to the information need. However, since `ss1[2]` contains no relevant information, `sec[1]` becomes *too broad*. The same is also true for `article[1]`. The set of relevant elements (or the full recall base) in this scenario consists of three elements: one *Just right* and two *Broad*.

*Scenario 2*: Assume that both `ss1[1]` and `ss1[2]` are relevant to an information need, and they also mainly just cover its aspects. The `sec[1]` element in this case contains two *Just right* children, which also makes it *Just right*. Indeed, the two `ss1` elements may cover two different aspects of the information need, or they may cover a single aspect from two different perspectives. Since the additional context provided by `sec[1]` is (arguably) more desirable than each of the two separate contexts of its children, both the `ss1` elements become *too narrow*. Also, since `sec[2]` contains no relevant information, `article[1]` becomes *too broad*. The full recall base in this scenario consists of four relevant elements: one *Just right*, one *Broad*, and two *Narrow*.

*Scenario 3*: Assume that the three elements, `ss1[1]`, `ss1[2]`, and `sec[2]`, are relevant to an information need, and all of them mainly just cover its aspects. The full recall base in this scenario consists of five relevant elements: one *Just right* and four *Narrow*, where `article[1]` is the only element that is *Just right*.

### *Exploring Aspects of XML Retrieval*

Different aspects of XML retrieval may be explored by using the new relevance definition.

One aspect would be to measure the XML retrieval effectiveness when only *Just right* elements are considered in the retrieval task. Note that in this case the full recall base consists of non-overlapping relevant elements, so there is no overlap problem during evaluation.

Another aspect would be to separately consider the *Broad* and the *Narrow* relevant elements in the recall base, and to measure the retrieval effectiveness against each of these elements. Indeed, different topics (or queries) require different granularity or relevant elements [15]. However, in both of these cases different techniques may be needed to deal with the overlap problem.

Previous work done by Voorhees in the field of Web retrieval confirms the hypothesis that different retrieval techniques need be used to retrieve highly relevant, rather than just any relevant, Web pages [21]. It may thus be worthwhile exploring whether, in the field of XML retrieval, different retrieval techniques would be needed to retrieve *Just right*, rather than any *Broad* or *Narrow*, relevant units of information.

### *Comparison with the INEX Relevance Definition*

Compared to the current INEX relevance definition, the new definition of relevance is much simpler. Indeed, instead of having a 10-point relevance scale that uses various combination of grades of the two INEX dimensions as values, the new relevance definition uses a four-point relevance scale with the following values: *Non-relevant*, *Narrow*, *Just right*, and *Broad*.

Also, more than one mappings may be possible between the INEX relevance definition and the new one. For example, a partial mapping of the new four-point relevance scale to the INEX 10-point relevance scale is as follows.

1. *Non-relevant* `<=> E=0, S=0 (E0S0)`
2. *Just right* `<=> E=3, S=3 (E3S3)`
3. *Broad* `<=> E=3, S<3 (E3S2, E3S1)`
4. *Narrow* `<=> E<3, S=3 (E2S3, E1S3)`

The above mapping is partial as it does not include the following four INEX relevance points: `E2S2`, `E2S1`, `E1S2`, and `E1S1`. One reason for this is that we choose only a highly relevant (`E3S3`) element on a path to represent a *Just right* element. From the properties of the new relevance definition (as outlined above), it follows that a *Broad* or a *Narrow* element could then be either *above* or *bellow* the *Just right* element, which limits the mapping choices. Another reason, however, stems from the fact that these four points of the relevance scale were not well perceived by both the assessor and the users. The latter may be the most probable cause for the observed inconsistencies regarding the *Specificity* dimension. Nevertheless, for the purposes of the evaluation of XML retrieval there is almost no need to modify some of the current INEX metrics in order to use the new relevance definition.

The new relevance definition could also easily be applied to the recent proposal of performing the assessor's relevance judgements at INEX 2005. This proposal is as follows: first, for a returned article the assessor will be asked to highlight all of the relevant content. Second, after the assessment tool automatically identifies the elements that enclose the highlighted content, the assessor will need to judge the level of *Exhaustivity* of these elements and of all their ancestors. Last, based on the highlighted text, the level of *Specificity*

will be computed automatically as a ratio of relevant to non-relevant information, however a mapping may be needed to get the four relevance grades for the *Specificity* dimension.

Although we agree that the above approach is very promising, it is still unclear whether keeping the current INEX relevance dimensions, along with their corresponding grades, would help reducing the cognitive load of the assessor (or the users) while performing the relevance judgements. The new relevance definition, on the other hand, is much simpler, and it also fits very nicely with the above proposal.

# 7. CONCLUSIONS

In this work, we have undertaken a detailed analysis of assessor's and users' behaviour in the context of XML retrieval. We have shown that the two relevance dimensions used by INEX, *Exhaustivity* and *Specificity*, are not orthogonal and are perceived as one dimension by users. By analysing the level of agreement between the assessor and the users, we also wanted to identify how both of them perceive the points of the INEX 10-point relevance scale; the results of our analysis show that the highest level of agreement is on the end points of the relevance scale, which means that a much simpler relevance scale would be a preferable choice for the field of XML retrieval. We have proposed a new definition of relevance to be used by INEX, and argued that its corresponding relevance scale is simpler and more comprehensive than the one currently used.

Our analysis also shows that, although the assessor handles the overlap problem differently than users, in the users' case there is almost no overlap between the elements judged as relevant. The latter confirms the hypothesis that users do not want to retrieve, and thus do not tolerate, redundant information.

We have not discussed how the overlap problem may be modelled by the new relevance definition. As argued previously, it may be possible to model the overlap problem by using a separate relevance dimension based on novel relevance, which can be integrated into the *Context* component of the unified relevance framework [13]. However, in this paper we do not pursue this discussion any further.

The observed retrieval behaviour of the assessors and users was based on two topics, each from a different topic category. We did not observe any notable differences among the above behaviours for the two topics. However, analysis of a greater number of topics is needed to confirm the significance of our findings. This will enable a comparison between the observed and the overall behaviour of the assessors and users, which will certainly establish the XML retrieval environment in a more consistent manner. We leave the activities related to this analysis for future work.

It is our hope that, by analysing the different aspects of the observed retrieval behaviour, the work presented in this paper will aid better understanding of the important issues surrounding INEX and the field of XML retrieval.

## Acknowledgements

We thank Saied Tahaghoghi and the anonymous reviewers for providing useful comments on earlier drafts of this paper.

# 8. REFERENCES

[1] J. Allan, R. Gupta, and V. Khandelwal. Temporal summaries of news topics. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 10–18, 2001.

[2] P. Borlund. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research*, 8(3):1–38, 2003. `http://informationr.net/ir/8-3/paper152.html`.

[3] J. Carbonell and J. Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336, 1998.

[4] K. Finesilver and J. Reid. User behaviour in the context of structured documents. In *Proceedings of the 25th European Conference on IR Research (ECIR), Pisa, Italy, April 2003*, pages 104–119, 2003.

[5] N. Fuhr, M. Lalmas, S. Malik, and Z. Szlavik, editors. *Advances in XML Information Retrieval: Third International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2004, Dagstuhl Castle, Germany, December 6-8, 2004, Revised Selected Papers*, volume 3493. Springer-Verlag GmbH, May 2005.

[6] N. Gövert, N. Fuhr, M. Abolhassani, and K. Großjohann. Content-oriented XML retrieval with HyREX. In *Proceedings of the First International Workshop of the INitiative of the Evaluation of XML Retrieval, INEX 2002, Dagstuhl Castle, Germany, December 8–11, 2002*, pages 26–32, 2003.

[7] K. Hatano, H. Kinutan, M. Watanabe, Y. Mori, M. Yoshikawa, and S. Uemura. Keyword-based XML fragment retrieval: Experimental evaluation based on INEX 2003 relevance assessments. In *Proceedings of the Second International Workshop of the INitiative of the Evaluation of XML Retrieval, INEX 2003, Dagstuhl Castle, Germany, December 15–17, 2003*, pages 81–88, 2004.

[8] G. Kazai, M. Lalmas, and A. P. de Vries. The overlap problem in content-oriented XML retrieval evaluation. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 72–79, 2004.

[9] G. Kazai, S. Masood, and M. Lalmas. A study of the assessment of relevance for the INEX 2002 test collection. In *Proceedings of the 26th European Conference on IR Research (ECIR), Sunderland, UK, April 2004*, pages 296–310, 2004.

[10] V. Lavrenko. A generative theory of relevance. *PhD dissertation*, University of Massachusetts, Amherst, MA, 2004. `http://ciir.cs.umass.edu/~lavrenko/thesis.pdf`.

[11] J. A. List and A. P. de Vries. XML-IR: Coverage as a part of relevance. In *Proceedings of the Dutch-Belgian IR Workshop (DIR), Leuven, Belgium, December 2002*, pages 7–12, 2002.

[12] S. Malik, M. Lalmas, and N. Fuhr. Overview of INEX 2004. In Fuhr et al. [5], pages 1–15.

[13] S. Mizzaro. Relevance: The whole history. *Journal of the American Society for Information Science and Technology*, 48(9):810–832, 1997.

[14] J. Pehcevski, J. A. Thom, S. M. M. Tahaghoghi, and A.-M. Vercoustre. Hybrid XML retrieval revisited. In Fuhr et al. [5], pages 153–167.

[15] J. Pehcevski, J. A. Thom, and A.-M. Vercoustre. Hybrid XML retrieval: Combining information retrieval and a native XML database. *Information Retrieval*, 8(4):571–600, 2005.

[16] B. Piwowarski and M. Lalmas. Providing consistent and exhaustive relevance assessments for XML retrieval evaluation. In *Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management (CIKM '04)*, pages 361–370, 2004.

[17] T. Saracevic. Relevance reconsidered. In *Proceedings of the Second Conference on Conceptions of Library and Information Science (COLIS), Copenhagen, Denmark*, pages 201–218, 1996.

[18] I. Soboroff and D. Harman. Overview of the TREC 2003 Novelty track. In *Proceedings of the Twelfth Text REtrieval Conference (TREC-12)*, pages 38–53, 2004.

[19] A. Tombros, B. Larsen, and S. Malik. The Interactive track at INEX 2004. In Fuhr et al. [5], pages 410–423.

[20] A. Tombros, I. Ruthven, and J. M. Jose. How users assess web pages for information seeking. *Journal of the American Society for Information Science and Technology*, 56(4):327–344, 2005.

[21] E. M. Voorhees. Evaluation by highly relevant documents. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82, 2001.

[22] C. Zhai. Risk minimization and language modeling in text retrieval. *PhD dissertation*, Carnegie Mellon University, Pittsburgh, PA, 2002. `http://www.cs.cmu.edu/~czhai/thesis.pdf`.